

EMNLP 2026

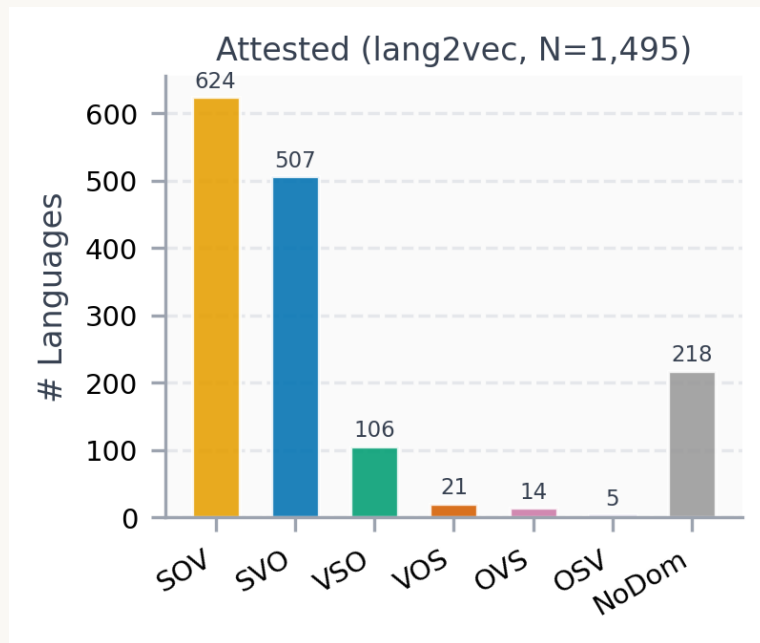
Left-Branching Transformers Excel at Right-Branching Languages

Data Shapes Word Order Preferences in Language Models

Varvara Arzt · Allan Hanbury · Terra Blevins

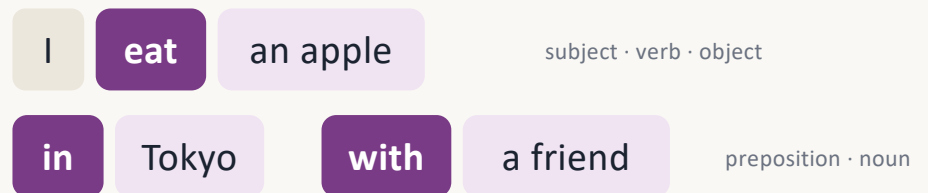
TU Wien · Northeastern University

Natural Languages Vary a Lot in their Word Order



ENGLISH · SVO

heads first → right-branching



JAPANESE · SOV

heads last ← left-branching



Do LMs Handle All Word Orders Equally Well?

RESEARCH QUESTIONS

RQ1

Do LMs have systematic word order preferences?

RQ2

If there are systematic preferences, are these architectural or data-driven?

RQ3

Do LM preferences mirror attested patterns?

RQ4

Do LM preferences on artificial languages match those on natural languages?

WHY BOTH?

Artificial languages only

Controlled, but no semantics, morphology or discourse:
findings may not transfer

Natural languages only

Realistic, but word order is confounded with everything else

→ **We do both: same architecture, matched training data scales**

Prior work studies one setting, with mixed results: White & Cotterell 2021; Kallini et al. 2024; Kuribayashi et al. 2024; El-Naggar et al. 2025; Ziv et al. 2026

Arzt, Hanbury & Blevins · EMNLP 2026

Experimental Setup

	Natural languages	Artificial languages
Languages	>70 typologically diverse languages Same language set at every data size	192 PCFG-generated languages 6 base orders $\times$ 2^5 internal constituent orderings
Models	Monolingual Goldfish (GPT-2 architecture) + multilingual mGPT, XGLM, BLOOM	GPT-2 small , trained from scratch Matched to Goldfish at 5 & 10 MB
Training data	Web text: word order + everything else 5 MB $\rightarrow$ 1 GB per language, byte-premium scaled	Identical data, only word order differs 5 & 10 MB
Evaluation	FLORES-200 & PUD BPEC (bits per English character)	Held-out sentences Perplexity

Shared across both settings: **GPT-2 architecture** · 5 & 10 MB of training data

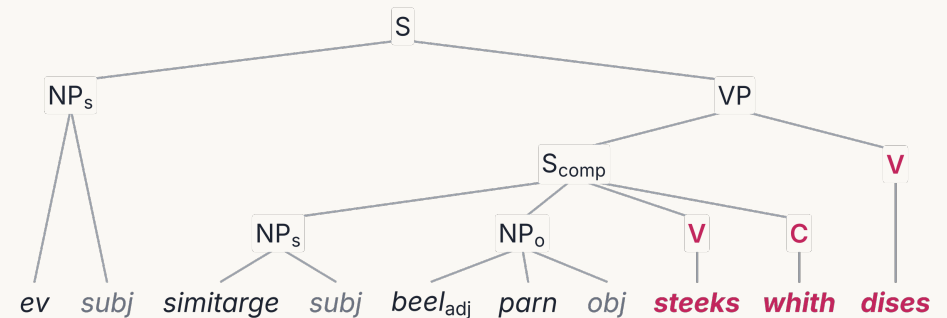
Building Artificial languages

Switch	L	R
1 Complement clause	S-comp V	V S-comp
2 Complementiser	S Comp	Comp S
3 Adposition	NP Postp	Prep NP
4 Adjective	Adj N	N Adj
5 Relative clause	VP Rel N	N Rel VP

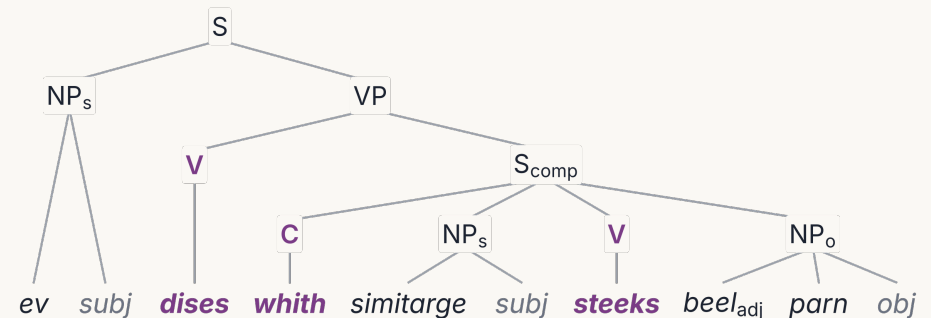
6 base orders × 2⁵ = 192 languages

English-like pseudowords; *subj* / *obj* = overt case markers.
Heads (verbs, complementiser) highlighted.

SOV-LLLLL ~ Japanese



SVO-RRRLR ~ English



Labeling Natural Languages the Same Way

ONE LABEL PER LANGUAGE

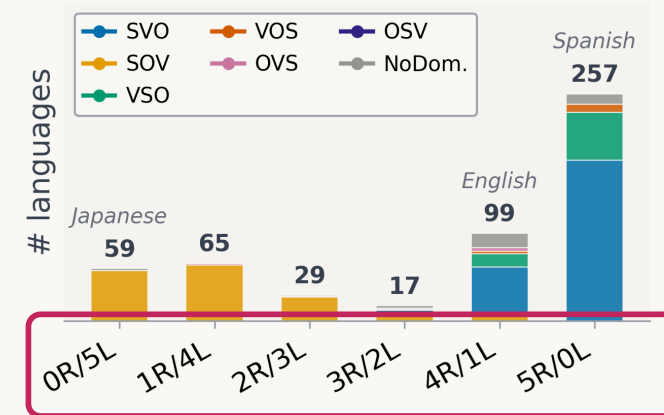
SV0- **R** **R** **R** **L** **R**

1 complement clause 2 complementiser 3 adposition 4 adjective 5 relative clause

base word order + 5 binary switches: **L** left- / **R** right-branching

Language	Label	Scale
English	SV0- RRRLR	4R/1L
French	SV0- RRRRR	5R/0L
Hindi	SOV- LRLLL	1R/4L
Japanese	SOV- LLLLL	0R/5L

Documented natural language distribution (N=526)



Scale = number of right-branching switches, 0–5: the x-axis of most of our results plots

Natural languages cluster at **both extremes**

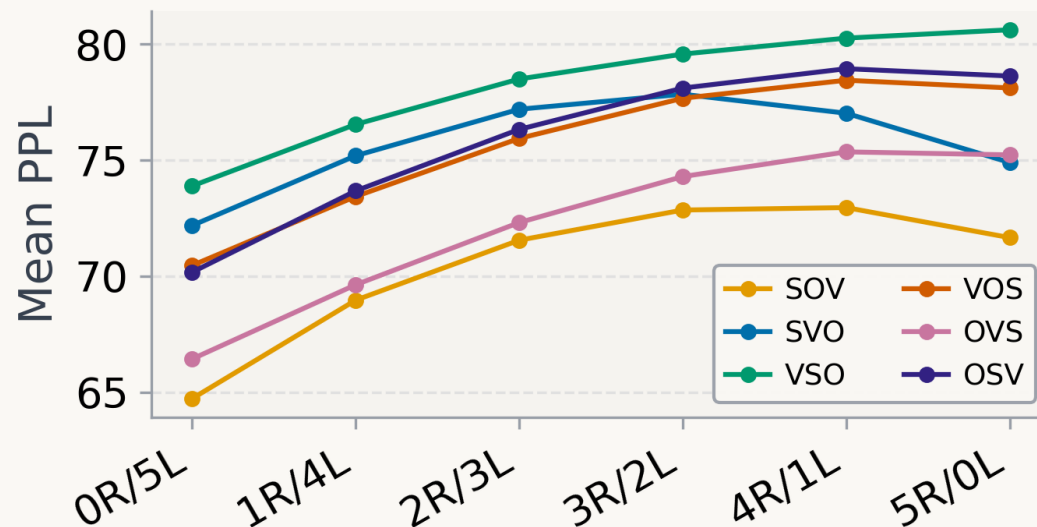
526 languages with binary values for all five switches

Labels from typological databases (WALS, Grambank, APiCS); we only use languages for which all five switches have binary values

LM Preferences in **Artificial** Languages

Left-branching is easiest for all six base word orders, **consistently at 5 and 10 MB**

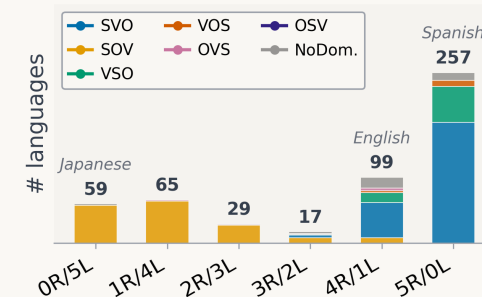
Artificial languages, 5 MB (GPT-2, N=192)



x-axis: number of right-branching switches (0R/5L = fully left-branching, e.g. SOV-LLLLL). GPT-2, 5 MB, mean test PPL over 10 runs (↓ easier).

Perplexity rises with each right-branching switch

Documented natural language distribution (N=526)



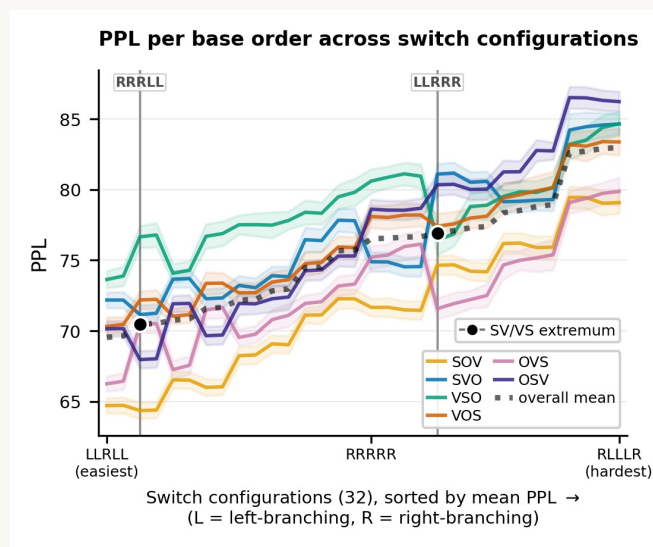
Natural languages: **SOV** mostly left-, **SVO** mostly right-branching. **LMs show no such pattern on artificial languages**

The easiest configuration for each base word order has **no natural counterpart**

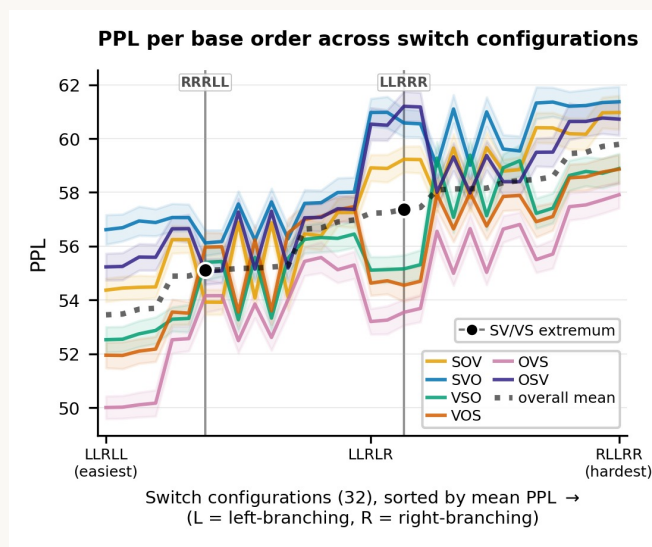
LM Preferences in Artificial Languages

Relative clauses drive left-branching; base word order rankings **flip** with data

GPT-2 · 5 MB



GPT-2 · 10 MB



Internal constituent order, not base word order, drives the spread: all six lines rise together

Relative clauses matter most:

prenominal ones are 4–7 PPL easier, against typology (postnominal ~4× more common)

Base word orders group by **SV vs VS**, not by **VO vs OV** as in natural languages

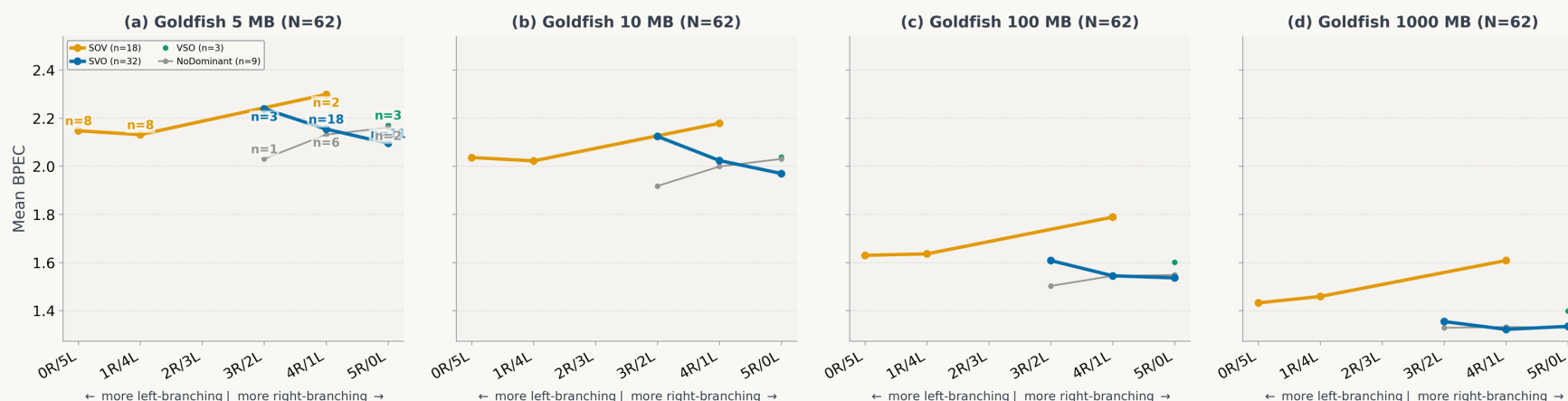
5 MB: SOV easiest, VSO hardest

10 MB: VS orders easiest, **SVO** hardest

All 32 switch configurations, sorted by mean PPL (left = easiest, L = left-, R = right-branching); one line per base word order. GPT-2, mean over 10 runs.

LM Preferences in Natural Languages

No preference at 5 MB. With more data, right-branching SVO becomes **easiest**.



SVO-SOV δ	0.00	-0.10	-0.60***	-0.76***
------------------	------	-------	----------	----------

Opposite of artificial languages: there, SVO is the **hardest** base word order at 10 MB; here, the **easiest** at scale

Same languages at every size: by 1 GB, SOV is $\approx 10\%$ harder to compress, though it is the most frequent order worldwide

Goldfish monolingual models, FLORES-200, mean BPEC ($\downarrow$ easier). δ = Cliff's δ (< 0 : SVO easier), *** $p < .001$; full table in paper Table 1.

What Drives the SVO Advantage?

Linear mixed-effects model: the growing SVO–SOV gap survives every control

Not script. 71% of SOV vs 19% of SVO FLORES-200 languages are non-alphabetic, but on alphabetic scripts only the gap still grows with data

Not morphological complexity. Its effect shrinks with data, so it cannot explain a gap that grows

Not training data composition. Web-crawl (OSCAR) vs. curated share of each language's data: no effect on the growth

Partly language resourcedness (≈20% of the effect): a proxy for data quality

50 SVO/SOV languages × 4 sizes; word order × size interaction with per-language random intercepts (paper Appendix Table 6)

Multilingual Models: the Same SVO Advantage

Model	SVO	SOV	NoDom	δ
BLOOM-7b1	2.81 (.206)	1.82 (.29)	—	+.25
XGLM-1.7B	1.41 (.88)	1.78 (5.12)	1.42 (.54)	-.53*
mGPT-13B	1.44 (.25)	1.68 (.18)	1.54 (.24)	-.44*

Median FLORES-200 BPEC (IQR); ↓ better, bold = best per model. δ = SVO–SOV Cliff's δ (< 0: SVO easier). * $p < .05$
Largest model per family; the same pattern holds at all sizes (560M–13B, backup).

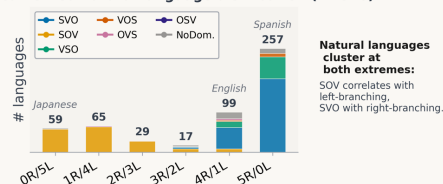
XGLM and mGPT favour SVO, though built to broaden language coverage

BLOOM: 21 of its 28 SVO languages used in pretraining are Niger-Congo, mostly very-low-resource, which inflates the SVO group's median BPEC

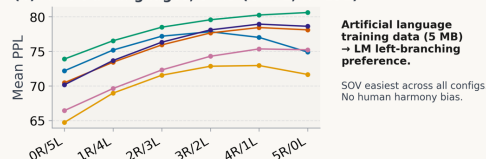
SVO languages get far more training data: e.g., median SVO/SOV ratio **5×** in XGLM

Takeaways

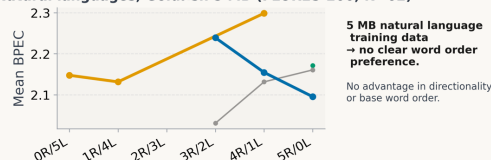
(a) Documented natural language distribution (N=526)



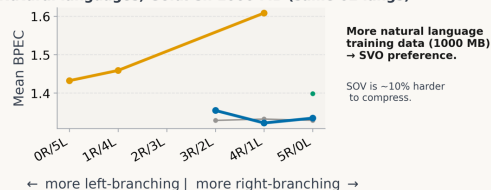
(b) Artificial languages, 5 MB (GPT-2, N=192)



(c) Natural languages, Goldfish 5 MB (FLORES-200, N=62)



(d) Natural languages, Goldfish 1000 MB (same 62 langs)



← more left-branching | more right-branching →

- 1 **Systematic, but not human-like: left-branching** on artificial languages regardless of base word order and against word order correlations in natural languages
- 2 **Data overrides the architectural bias:** on natural languages at scale, better-resourced **SVO** languages become easiest, in mono- and multilingual models
- 3 **LM preferences on artificial and natural languages go in opposite directions at scale:** **SVO** is hardest on artificial languages, easiest on natural ones

Even monolingual models trained on just 1 GB develop the SVO preference; frontier models, with far larger resource gaps between languages, likely amplify it.

A precondition for *silent typological homogenisation*: flexible word order languages drifting towards the dominant ones.



Paper & code
kleines-gespenst.github.io